Clustering Approach for Partitioning Directional Data in Earth and Space Sciences

Christian D. Klose[1] & K. Obermayer[2]

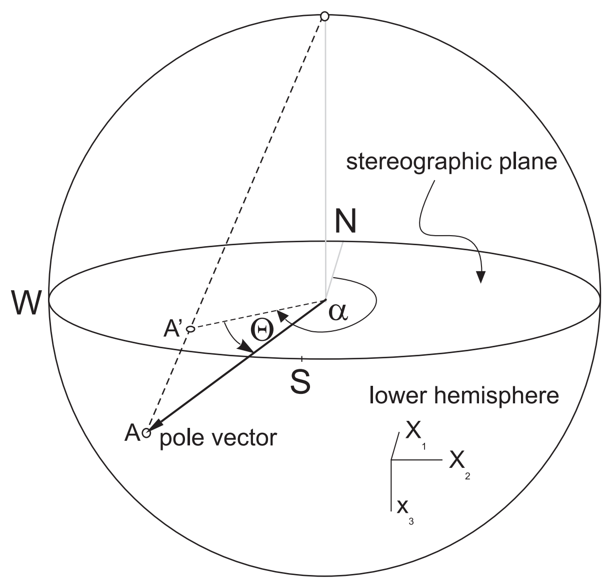
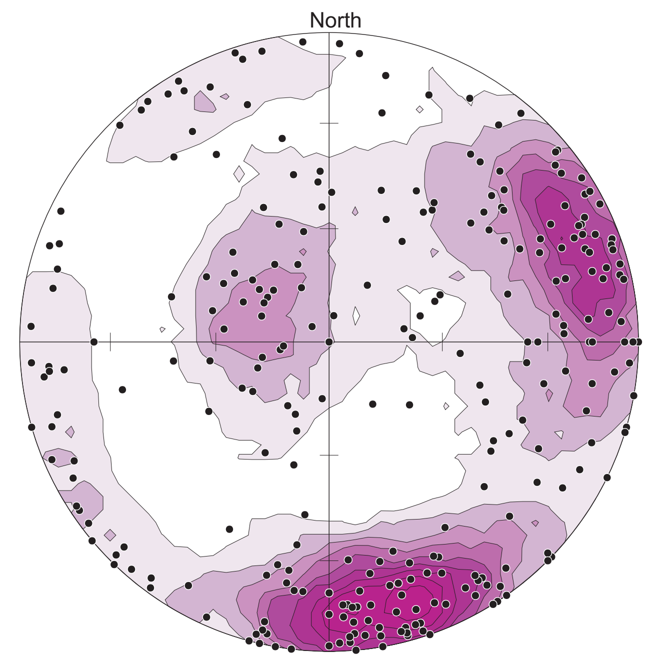
[1]Think GeoHazards, NY

[2]Berlin Institute of Technology, Germany

Introduction

- Clustering of bi/directional data is a fundamental problem in Earth and Space sciences,
- Counting methods in stereographic plots (Schmidt 1925; Shanley and Mahtab, 1976; Wallbrecher, 1978),
- Methods based on an iterative, stochastic reassignment of orientation vectors (Fisher 1987, Dershowitz et al. 1996),
- Methods based on fuzzy sets and on a similarity measure $d^2(\vec{x}, \vec{w}) = 1 - (\vec{x}^T \vec{w})^2$ (Hammah and Curran, 1998),

Introduction

An Orientation Vector	Stereographic Plot
	

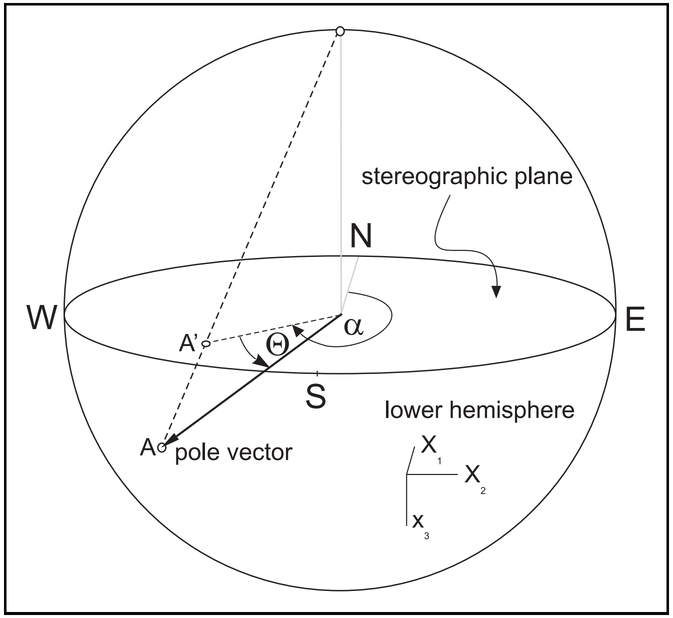
Sampling biases!!!

Motivation

- Cluster "pole vectors" $\vec{\Theta} = (\alpha, \theta)^T$
- Orientation $\vec{\Theta}^A = (\alpha^A, \theta^A)^T$ of a pole vector A, with $0° \leq \alpha \leq 360°$ and $0° \leq \theta \leq 90°$
- $\vec{\Theta}^A$ can be described by its Cartesian coordinates $\vec{x}^A = (x_1, x_2, x_3)^T$ as well, where

$$\begin{aligned} x_1 &= cos(\alpha)\, cos(\theta) && \text{North direction} \\ x_2 &= sin(\alpha)\, cos(\theta) && \text{East direction} \\ x_3 &= sin(\theta) && \text{downward.} \end{aligned} \tag{1}$$

Motivation



- We introduce a clustering method which is based on vector quantization (Gray 1984)
- Klose et al. (2005) A new clustering approach for partitioning directional data, IJRMMS.

The Clustering Method

- Assignment of pole vectors $\vec{x}_k$ to a partition

$$m_{lk} = \begin{cases} 1, & \text{if data point } k \text{ belongs to cluster } l \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

- Average dissimilarity between the data points and pole vectors

$$E = \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{M} m_{lk}\, d(\vec{x}_k, \vec{w}_l), \tag{3}$$

- Optimal partition by minimizing the cost function E, i.e.

$$E \stackrel{!}{=} \min_{\{m_{lk}\},\{\vec{w}_l\}} \tag{4}$$

The Clustering Method

- Minimization is performed iteratively in two steps.
- Step 1: cost function E is minimized with respect to $\{m_{lk}\}$

$$m_{lk} = \begin{cases} 1, & \text{if} \quad l = \arg\min_q \, d(\vec{x}_k, \vec{w}_q) \\ 0, & \text{else.} \end{cases} \tag{5}$$

- Step 2: E is minimized with respect to $\vec{\Theta}_l = (\alpha_l, \theta_l)^T$ which describe the average pole vectors $\vec{w}_l$:

$$\frac{\partial E}{\partial \vec{\Theta}_l} = \vec{0}, \tag{6}$$

The Clustering Method

BEGIN Loop

Select a data point $\vec{x}_k$.

Assign data point $\vec{x}_k$ to cluster l by:

$$l = \arg\min_q d(\vec{x}_k, \vec{w}_q) \quad (7)$$

Change average pole vector of this cluster by:

$$\Delta\vec{\Theta}_l = -\gamma \frac{\partial d(\vec{x}_k, \vec{w}_l(\vec{\Theta}_l))}{\partial \vec{\Theta}_l} \quad (8)$$

END Loop

The Distance Measure

- Distance measure $d(\vec{x}, \vec{w})$ must satisfy the following conditions
  1. $d(\vec{x}, \vec{w}) = \min \Leftrightarrow \vec{x}$ and $\vec{w}$ are equally directed parallel vectors, i.e. $\vec{x}^T \vec{w} = 1$.
  2. $d(\vec{x}, \vec{w}) = \max \Leftrightarrow \vec{x}$ and $\vec{w}$ are orthogonal vectors, i.e. $\vec{x}^T \vec{w} = 0$.
  3. $d(\vec{x}, \vec{w}_1) = d(\vec{x}, \vec{w}_2)$ if $\vec{w}_1$ and $\vec{w}_2$ are antiparallel vectors, i.e. $\vec{w}_2 = -\vec{w}_1$.
- arc-length between the projection points

$$d(\vec{x}, \vec{w}) = \arccos\left(\,|\,\vec{x}^T \vec{w}\,|\,\right), \tag{9}$$

The (online) Algorithm

Initialize: Pole vectors $\alpha_q(0),\ \theta_q(0),\ \forall\ q = 1, \ldots, M$, annealing schedule (learning rate $\gamma(t)$, maximum number t_F of iterations).

Set: Iteration number $t = 0$.

Compute: $\vec{w}_q(t) = \vec{w}_q(\alpha_q(t), \theta_q(t))^T$

The (online) Algorithm

Repeat

1. Draw $\vec{x}_k$ randomly from the data set.
2. Compute $d(\vec{x}_k, \vec{w}_q(t)) = \arccos |\, \vec{x}_k^T \vec{w}_q(t) \,|$ for all $q = 1, \ldots, M$.
3. Find index $l = \arg\min_q d(\vec{x}_k, \vec{w}_q(t))$ of pole vector $\vec{w}_l(t)$ closest to $\vec{x}_k$.
4. Compute the parameters $\alpha_l(t+1)$ and $\theta_l(t+1)$
5. Compute the pole vector $\vec{w}_l(t+1) = \vec{w}_l(\alpha_l(t+1), \theta_l(t+1))$
6. Compute new learning rate $\gamma(t+1) = \frac{\gamma(t)\gamma(t_F)}{\gamma(t_F)+t}$.
7. $t \leftarrow t+1$.

Until: $t > t_F$.

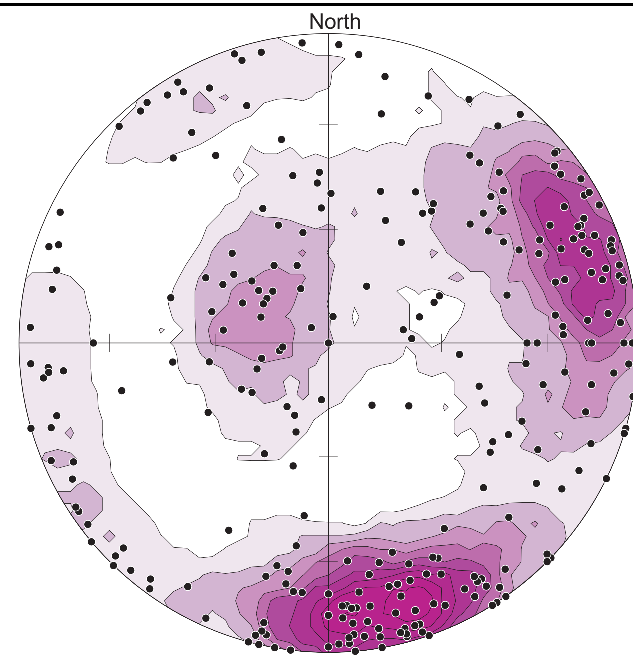
The (online) Algorithm

Project all $\vec{w}_q$, $q = 1, \ldots, M$ to the lower hemisphere (as defined):

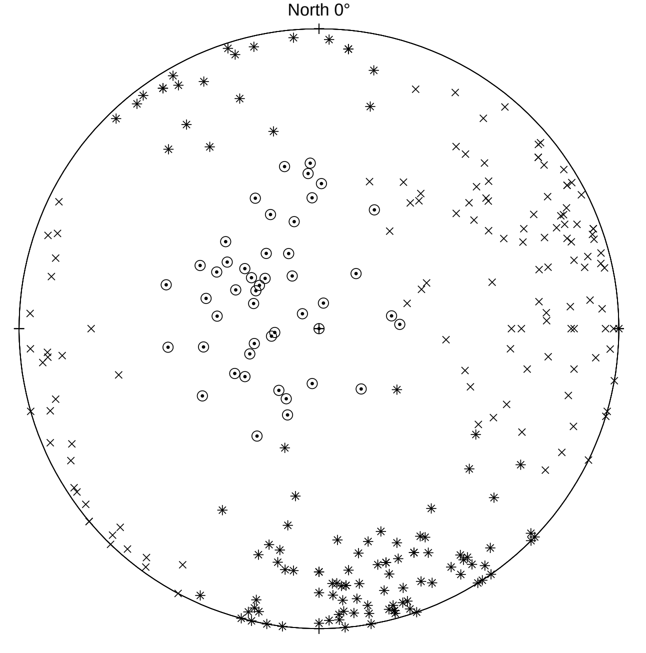
If the third component of the pole vectors $(\vec{w}_q)_3 > 0$, then

$$
\begin{aligned}
\vec{w}_q &= -\vec{w}_q, \\
\theta_q &= -\theta_q, \\
\alpha_q &= \begin{cases} \alpha_q + \pi & \text{if} \quad \alpha_q < \pi \\ 2\pi - \alpha_q & \text{if} \quad \alpha_q \geq \pi. \end{cases}
\end{aligned}
$$

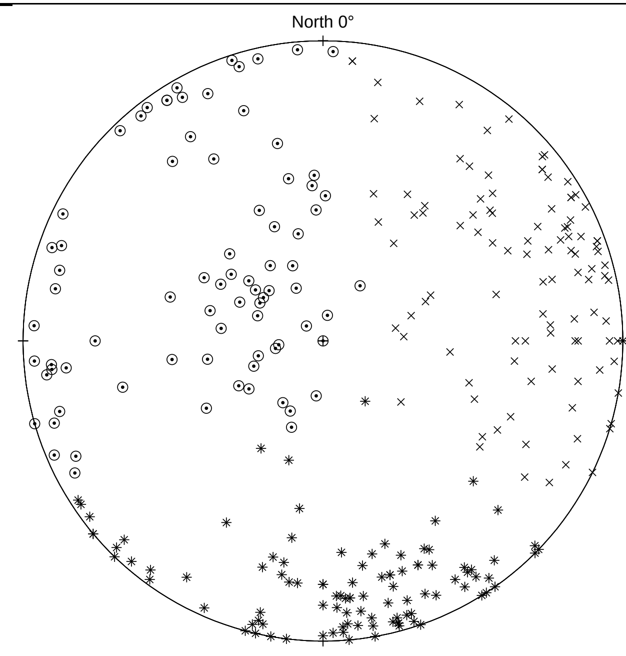
Application

Cluster	density plot
1 ($\times$)	72/14
2 ($\ast$)	163/14
3 ($\odot$)	303/81
	

Application

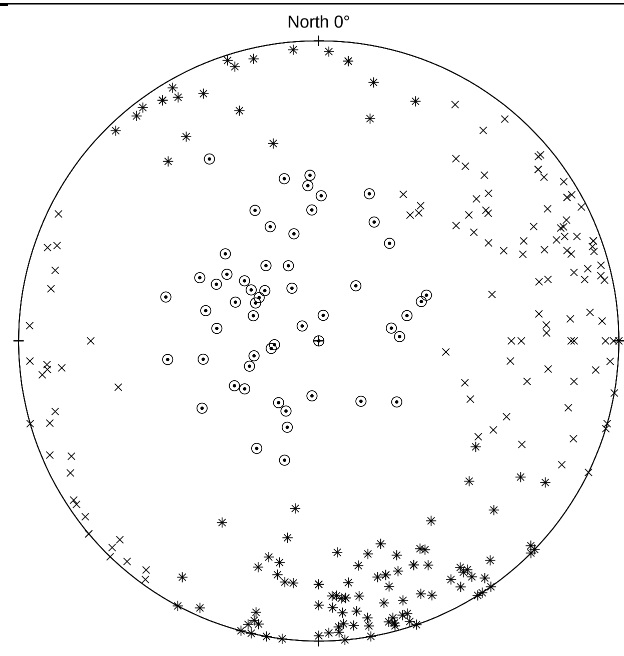
Cluster	Shanley & Mahtab
1 ($\times$)	72/14
2 ($*$)	163/14
3 ($\odot$)	303/81
	

Application

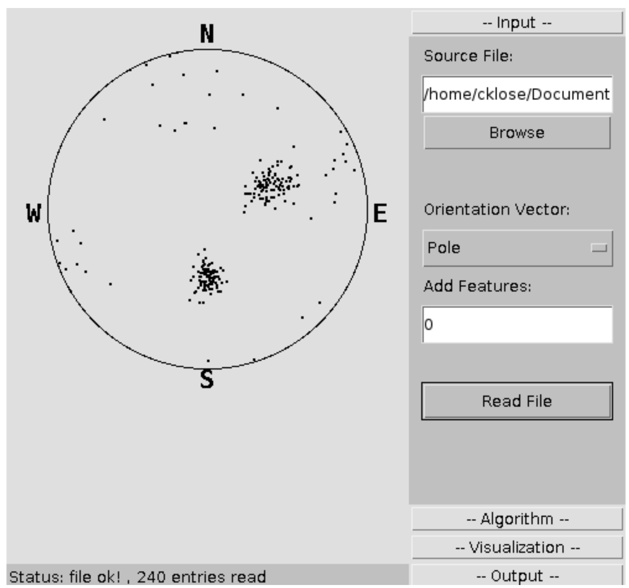
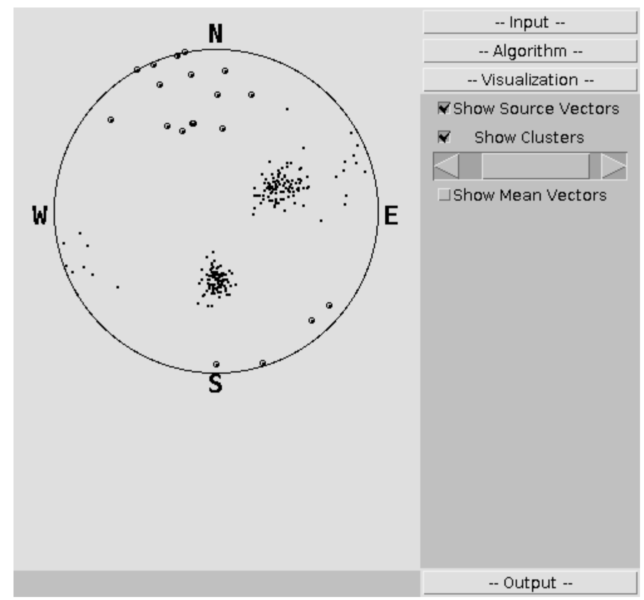
Cluster	Pecher
1 ($\times$)	71/24 (26%)
2 ($\ast$)	175/14 (21%)
3 ($\odot$)	299/46 (11%)
	

Application

Cluster	new clustering method
1 ($\times$)	68/15 (7%)
2 ($\ast$)	171/10 (3%)
3 ($\odot$)	310/73 (0%)



Application - Software App

Input	Output
	

URL: http://www.thinkgeohazards.com/index.TGH.html

Conclusion

- Partitioning directional data into disjoint isotropic clusters,
- Analysis of their average orientation,
- This new method is self-consistent (EM steps, same cost function),
- This method does not require special preprocessing,
- Ongoing research on probabilistic assignments (soft-clustering) and additional features.

Next Steps, e.g., Magnetic Data or Weather Data

Northern Hemisphere	Southern Hemisphere